AD-A101 734   PENNSYLVANIA STATE UNIV UNIVERSITY PARK DEPT OF STAT--ETC F/G 12/1
INFERENCE BASED ON SIMPLE STEP STATISTICS FOR THE LOCATION MODE--ETC(U)
JUL 81 E P MARKOWSKI, T P HETTMANSPERGER   N00014-80-C-0741
UNCLASSIFIED TR-37   NL

1 OF 1
AD
A101 734

END
DATE
FILMED
8-81
DTIC

LEVEL II

(12)

The Pennsylvania State University

Department of Statistics

University Park, Pennsylvania

LEVEL II

DEPARTMENT OF STATISTICS
The Pennsylvania State University
University Park, PA 16802, U.S.A.

TECHNICAL REPORTS AND PREPRINTS
Number 37: July 1981

INFERENCE BASED ON SIMPLE STEP STATISTICS FOR THE LOCATION MODEL

Edward P. Markowski
Old Dominion University

Thomas P. Hettmansperger*
The Pennsylvania State University

Accession For	
NTIS GRA&I	X
DTIC TAB	☐
Unannounced	☐
Justification	
By	
Distribution/	
Availability Codes	
Dist	Avail and/or Special
A	

*Research partially supported by ONR contract N00014-80-C-0741

1. Introduction

In this paper, we will be concerned with location models and, in particular, the one sample location model. We wish to make inferences about θ, the center of symmetry of a continuous, symmetric population with c.d.f. F, based on a random sample of size n. Many authors have considered this problem, although many place more restrictive assumptions on F, such as normality. If these additional assumptions are questionable, practitioners often seek procedures which remain valid when no specific form for the underlying distribution is assumed. For testing hypotheses about θ, the Wilcoxon and sign tests have emerged as two of the more popular procedures. In addition, estimation procedures, both point and interval, based on these rank statistics have been developed. However, the sign test, while optimal if F is the double exponential c.d.f., is quite inefficient at the normal model. Conversely, the Wilcoxon test, which is optimal if F is the logistic c.d.f., is less efficient for heavier tailed distributions. Also, the estimation techniques derived from the Wilcoxon statistic (Lehmann (1975), Chapter 4, Section 3) are computationally difficult for moderate or large samples (Johnson and Ryan (1978)).

Thus, we propose a class of rank statistics "between" the sign and Wilcoxon statistics. The corresponding tests and estimation procedures will retain much of the simplicity of those based on the sign statistic but with improved efficiency (especially near the normal model). In addition, all the procedures have excellent robustness properties.

2. Testing

Let $X_1 \leq \ldots \leq X_N$ be the ordered values of a random sample from F. Let $0 \leq a(1) \leq a(2) \leq \ldots \leq a(N)$ be a given set of scores. Then $W_N(\theta_o) = \Sigma a(i)V_i(\theta_o)$ is a linear signed rank statistic for testing H_o: $\theta = \theta_o$ where $V_i(\theta_o) = 1$ if $X_{d_i} - \theta_o > 0$ and 0 otherwise and $|X_{d_1} - \theta_o| \leq \ldots \leq |X_{d_N} - \theta_o|$. In this case, i is the rank of $|X_{d_i} - \theta_o|$ among the absolute values, i = 1, ..., n. Notice that a(i) = i yields the Wilcoxon statistic and a(i) = 1 yields the sign statistic. Our approach is to select a set of scores "between" the sign and Wilcoxon scores. Let

$$a_k(i) = \begin{cases} 0 & i = 0, \ldots, b_1 \\ 1 & i = b_1 + 1, \ldots, b_2 \\ 2 & i = b_2 + 1, \ldots, b_3 \\ \vdots & \vdots \\ k & i = b_k + 1, \ldots, N \end{cases} \tag{2.1}$$

where $0 = \gamma_o \leq \gamma_1 < \gamma_2 < \ldots < \gamma_k \leq \gamma_{k+1} = 1$, $b_t = [N\gamma_t]$ $t = 0, \ldots, k + 1$ and $[\cdot]$ is the greatest integer function. Let $T_{N,k}(\theta) = \Sigma a_k(i)V_i(\theta)$. Then $T_{N,k}$ is called the k-step statistic.

Noether (1973) studied the 1-step statistic with particular emphasis on interval estimation based on $T_{N,1}$. Notice that if $\gamma_1 = 0$, $T_{N,1}$ is the sign statistic. Policello and Hettmansperger (1976) discuss the 1-step statistic with regard to adaptive methods of estimation of θ. Notice that the freedom to vary $\gamma_1, \ldots, \gamma_k$ makes this a rich family of statistics and hence some member of the family should be suitable for testing H_o over a wide range of symmetric underlying distributions.

To test H_o: $\theta = \theta_o$ vs. H_a: $\theta > \theta_o$, we will reject H_o if $T_{N,k}(\theta_o) > c$. However, we will consider only the case $\theta_o = 0$ since otherwise we can subtract θ_o from each sample observation. Henceforth, we denote $T_{N,k}(0)$ by $T_{N,k}$. Next, we consider the null distribution of $T_{N,k}$.

Theorem 1. If $\theta = 0$, then $T_{N,k} = \sum_{i=1}^{k} iW_i$ where W_i is distributed as binomial with parameters $b_{i+1} - b_i$ and 1/2, and $W_1, \ldots, W_k$ are independent.

This result follows directly from Lemma 10.1.11 of Randles and Wolfe (1979).

Corrollary. If $\theta = 0$, then

$$E(T_{N,k}) = \sum_{i=1}^{k} i(b_{i+1} - b_i)/2, \quad Var(T_{N,k}) = \sum_{i=1}^{k} i^2(b_{i+1} - b_i)/4$$

Notice that $T_{N,1}$ follows a binomial distribution with parameters $N - b_1$ and 1/2 and that $T_{N,2}$ is a weighted combination of two independent binomial random variables. We conclude discussion of the null distribution of $T_{N,k}$ with the following result on the asymptotic distribution of $T_{N,k}$.

Theorem 2. Under H_o: $\theta = 0$, for $k \geq 1$ fixed, the limiting distribution (as $N \to \infty$) of $(T_{N,k} - E(T_{N,k}))/(Var(T_{N,k}))^{1/2}$ is N(0, 1).

Proof. Theorem 1 exhibits $T_{N,k}$ as a linear combination of a fixed number of independent binomial random variables. Since each of these converges to a normal distribution, the result follows.

Example. Thirty observations were generated from a normal distribution with $\mu = 0$, $\sigma = 4$. The ordered data (rounded to two decimal places) follows.

-10.33	-2.83	0.01	2.30	4.28
-6.16	-1.36	0.05	2.38	4.30
-6.03	-1.30	0.31	2.97	4.92
-5.80	-0.93	1.25	3.47	6.19
-5.10	-0.53	1.81	3.67	6.32
-3.71	-0.52	2.24	4.08	7.96

We will compute the 2-step statistic with steps at $\gamma_1 = .2$, $\gamma_2 = .6$. The reason for this choice will be apparent later in this section. Since $N = 30$, $b_1 = [30(.2)] = 6$, $b_2 = 18$. Thus, $T_{N,2}$ ignores the six observations with smallest absolute value, counts the number of positive observations with absolute values ranking from 7 to 18, and adds to this sum twice the number of positive observations with absolute values ranking above 18. Thus, $T_{N,2} = 8 + 2(7) = 22$. For testing H_0: $\theta = 0$ vs. H_a: $\theta > 0$, we have, correcting for continuity, $Z = (T_{N,k} - .5 - E(T_{N,2}))/(\mathrm{Var}T_{N,2})^{1/2}$ and so $Z = (22 - .5 - 18)/\sqrt{15} = .90$ or $p = .1841$.

The exact distribution of $T_{N,k}$ under H_a is quite complex and will not be discussed. We mention only that Markowski (1980) has shown that the exact distribution of $T_{N,1}$ depends only on $P(X_j + X_{j+b_1} > 0)$ $j = 1, \ldots, N - b_1$, but the form of the dependence is quite complex.

In order to discuss asymptotic properties of the inference procedures studied, it is necessary to discuss some aspects of the asymptotic distribution of $T_{N,k}$ under H_a. First, Markowski (1980) has shown that $T_{N,k}$ is asymptotically normally distributed under a sequence of contiguous location alternatives. The proof involves a modification of a technique employed by Albers, Bickel, and van Zwet (1976). Next, when the limit exists, define the Pitman efficacy of a statistic R_N by $C(R_N) = \lim \frac{d}{d\theta} E_\theta(R_N)\big|_{\theta=0} / (N\ \mathrm{Var}_0\ R_N)^{1/2}$ where the variance is computed under the null hypothesis and the mean is computed under the alternative. Under regularity conditions, the asymptotic performance of two test statistics can be compared by considering the ratio of their squared efficacies. See Lehmann (1975, Appendix, Section 6) for a complete discussion.

The efficacy of the k-step statistic can be computed directly from the following.

<u>Theorem 3</u>. If F has finite Fisher's information, and the density, f, of F is differentiable, then

$$C^2(T_{N,k}) = \frac{4[\int_0^\infty J^+(2F(x) - 1)\ f'(x)dx]^2}{\int_0^1 J^+(u)du} \tag{2.2}$$

where J^+, the limiting scores function, is given by

$$J^+(u,\ k) = \begin{cases} 0 & 0 \le u \le \gamma_1 \\ 1 & \gamma_1 < u \le \gamma_2 \\ \cdot & \cdot \\ \cdot & \cdot \\ \cdot & \cdot \\ k & \gamma_k < u \le 1 \end{cases} \tag{2.3}$$

<u>Proof</u>. Begin with 10.2.11, p. 338 of Randles and Wolfe (1979) and apply an integration by parts to yield the result.

<u>Corollary</u>. Under the conditions of Theorem 3, the squared efficacies of the one and two step statistics are:

$$C^2(T_{N,1}) = \frac{4f^2\{F^{-1}[(1 + \gamma_1)/2]\}}{1 - \gamma_1} \tag{2.4}$$

and

$$C^2(T_{N,2}) = \frac{[f\{F^{-1}[(1 + \gamma_1)/2]\} + f\{F^{-1}[(1 + \gamma_2)/2]\}]^2}{1 - \gamma_2 + (\gamma_2 - \gamma_1)/4} \tag{2.5}$$

Using (2.4), we determine in Table 1 that member of the class of one-step statistics which maximizes the efficacy for the following underlying distributions:

A_1: Normal, $f[F^{-1}(x)] = (2\pi)^{-1/2}\{\exp[(F^{-1}(x))^2/2]\}$,

A_2: Logistic, $f[F^{-1}(x)] = x(1 - x)$,

A_3: Contaminated Normal, $\varepsilon = .05$, $\sigma^2 = 9$, $f(x) = (1 - \varepsilon)\phi(x) + \frac{\varepsilon}{\sigma}\phi(x/\sigma)$ where $\phi(\cdot)$ is the standard normal density function,

A_4: Double Exponential, $f[F^{-1}(x)] = x$ if $0 \leq x \leq .5$ and $1 - x$ if $.5 < x \leq 1$.

- Table 1 about here -

Table 1 indicates that for the 1-step statistic, the optimal choice of γ_1 decreases as the tails of the underlying model become heavier. Also, we see that despite the simplicity of the 1-step statistics, the best 1-step statistic is quite efficient compared to the optimal choice in the class of possible statistics. In addition, we have included a comparison of the best 1-step statistic as well as the 1-step with $\gamma_1 = 1/3$ as opposed to the sign statistic. These latter two comparisons were first discussed by Noether (1973) in a somewhat different setting. Notice that the statistic with $\gamma_1 = 1/3$ seems to be an improvement over the sign statistic since its efficacy is more stable and improves significantly over the sign statistic at the normal model.

- Table 2 about here -

Table 2 illustrates the effect on $C^2(T_{N,2})$ caused by changes in γ_1 and γ_2. In addition, the optimal choice of (γ_1, γ_2) is noted for each underlying model. Notice that the efficacy changes only slightly near the optimum value so that a choice of (γ_1, γ_2) in a neighborhood of the optimal value will still yield nearly highest efficiency. Of course, typically the underlying distribution may not be known and so a test with good efficiency over all distributions considered would be desirable. From Table 2, we see that the choice $\gamma_1 = .2$, $\gamma_2 = .6$ has such properties. In Table 3, we compare this 2-step statistic with the sign, Wilcoxon, and t statistics as well as the likelihood ratio statistic for the particular underlying distribution. We see that $T_{N,2}$ improves upon the sign statistic while retaining much of its simplicity. In addition, it compares quite favorably with the t-statistic at the normal model ($e = .89$) while being superior to the t-statistic at the other distributions considered. We see also that this 2-step statistic has properties very similar to that of the Wilcoxon statistic.

- Table 3 about here -

Thus, if one step is used, we recommend $\gamma_1 = 1/3$ and if two steps are used take $\gamma_1 = .2$, $\gamma_2 = .6$. These recommendations are based on the efficiency calculations of Table 3. In the remainder of the paper, we will investigate the estimation procedures associated with these tests and study their robustness properties.

3. Point Estimation

Hodges and Lehmann (1963) illustrate a method for deriving point estimators from rank statistics such as $T_{N,k}$. We denote by $\hat{\theta}^k(\gamma_1, \ldots, \gamma_k)$ or simply by $\hat{\theta}^k$ the Hodges Lehmann estimator (H-L) of θ derived from $T_{N,k}$. Bauer (1972) has shown that the linear rank statistics $W_N(\theta) = \Sigma a(i)V_i(\theta)$ as defined in Section 2 are non-increasing step functions which can have steps only at the Walsh averages, $\{(x_i + x_j)/2 : 1 \le i \le j \le N\}$. The sign statistic has steps only at $\{x_i : 1 \le i \le N\}$ while the Wilcoxon has steps at each of the Walsh averages. Following the method of Bauer (1972), we find that $T_{N,k}$ has steps of size 1 at the averages in $D_k = \{(x_i + x_j)/2 : j - i = b_t,\ t = 1, \ldots, k\}$ and hence $\hat{\theta}^k = \text{med } D_k$. Now for $t = 1, \ldots, k$ define $D_{k,t} = \{(x_i + x_{b_t+i})/2 : i = 1, \ldots, N - b_t\}$, the Walsh averages corresponding to the t'th step of $T_{N,k}$. Then $D_k = \bigcup_t D_{k,t}$ and so $\hat{\theta}^k = \text{med } (\bigcup_t D_{k,t})$. The advantage of this representation is that the Walsh averages in $D_{k,t}$ are a priori ordered. This means that for small k, finding the median $\hat{\theta}^k$ may not be difficult. For example,

$$\hat{\theta}^2 = \begin{cases} \{X_{(N - b_1 + 1)/2} + X_{(N + b_1 + 1)/2}\} /2 \text{ if } N - b_1 \text{ odd} \\ \{X_{(N - b_1)/2} + X_{(N + b_1)/2} + X_{(N - b_1 + 2)/2} + \\ \quad X_{(N - b_2 + 2)/2}\}/4 \text{ if } N - b_1 \text{ even.} \end{cases} \tag{3.1}$$

Thus, $\hat{\theta}^2$ can be computed directly from the ordered sample. As another example, we use the ordered structure of the sets D_{31} and D_{32} to exhibit a simple algorithm for computing $\hat{\theta}^3$. Our approach is to begin with medians of each of the sets D_{31} and D_{32} and then by a series of comparisons of Walsh averages, determine $\hat{\theta}^3$.

Let $m_{i,j} = (x_i + x_j)/2$ and $B(a, b) = m_{a,\ b_1+a} - m_{b,\ b_2+b}$

$a = 1, \ldots, N - b_1 \qquad b = 1, \ldots, N - b_2$

Algorithm

Case 1. $N - b_1$ even, $N - b_2$ odd.

First consider $B((N - b_1)/2, (N - b_2 + 1)/2)$ and suppose for now that it is greater than 0. Notice that this implies $m_{(N-b_2)/2,\ (N+b_2)/2} \leq \hat{\theta} \leq m_{(N-b_1)/2,\ (N+b_1)/2}$. Next, compute $B((N - b_1)/2, (N - b_2 + 1)/2 + 1)$. If $B((N - b_1)/2, (N - b_2 + 1)/2 + 1) < 0$, then $\hat{\theta} = m_{(N-b_1)/2,\ (N+b_1)/2}$. If $B((N - b_1)/2, (N - b_2 + 1)/2 + 1) > 0$, compute $B((N - b_1)/2 - 1, (N - b_2 + 1)/2 + 1)$. If $B((N - b_1)/2 - 1, (N - b_2 + 1)/2 + 1) > 0$, compute $B((N - b_1)/2 - 2, (N - b_2 + 1)/2 + 2)$. Continue until $B((N - b_1)/2 - i, (N - b_2 + 1)/2 + i) > 0$ and $B((N - b_1)/2 - (i + 1), (N - b_2 + 1)/2 + (i + 1)) < 0$. Then if $B((N - b_1)/2 - i, (N - b_2 + 1)/2 + (i + 1)) < 0$, $\hat{\theta} = m_{(N-b_1)/2-i,\ (N+b_1)/2-i}$ and if $B((N - b_1)/2 - i, (N - b_2 + 1)/2 + (i + 1)$ > 0, $\hat{\theta} = m_{(N-b_2+1)/2+(i+1),\ (N+b_2+1)/2+(i+1)}$. Recall that we assumed $B((N - b_1)/2, (N - b_2 + 1)/2) > 0$. If not, then proceed in a completely analogous fashion except increase the first component and decrease the second until $B((N - b_1)/2 + i, (N - b_2 + 1)/2 - i) > 0$.

Case 2. $N - b_1$ odd, $N - b_2$ even.

Proceed as in Case 1 but begin with $B((N - b_1 + 1)/2, (N - b_2)/2)$.

Case 3. $N - b_1$ odd, $N - b_2$ odd.

Begin with $B((N - b_1 + 1)/2, (N - b_2 + 1)/2)$. Proceed as in Case 1 reaching the point where (e.g.) $B((N - b_1 + 1)/2 + i, (N - b_2 + 1)/2 - i) < 0$ and $B((N - b_1 + 1)/2 + (i + 1), (N + b_2 + 1)/2 - (i + 1)) > 0$. If

$$B((N - b_1 + 1)/2 + (i + 1),\ (N - b_2 + 1)/2 - i) > 0 \qquad (3.2)$$

$$B((N - b_1 + 1)/2 + i,\ (N - b_2 + 1)/2 - (i + 1) < 0, \qquad (3.3)$$

then $\hat{\theta} = \{m_{(N-b_1+1)/2+(i+1),\ (N+b_1+1)/2+(i+1)} + m_{(N-b_2+1)/2-(i+1),\ (N+b_2+1)/2-(i+1)}\}/2$. If the signs of the inequalities (3.2) and (3.3) are other than those listed, then once again $\hat{\theta}$ can be determined in an analogous fashion.

Case 4. $N - b_1$ even, $N - b_2$ even.

Begin with $B((N - b_1)/2, (N - b_2)/2)$ and proceed as in Case 1. Then make comparisons as in (3.2) and (3.3). $\hat{\theta}$ is determined by the direction of these inequalities. Notice that it may happen that the final stage cannot be reached because the second component can no longer be increased (or decreased). At this point, $\hat{\theta}$ can be computed immediately since in this case:

1) $\hat{\theta}$ must come from the set D_{31}.

2) The position of the largest element of D_{32} in the ordering of D_{31} will have been determined by previous steps.

Next, we illustrate the use of the algorithm using the data of the example in Section 2. We will compute $\hat{\theta}(.2,.6)$ for this data. Now $N = 30$, $b_1 = 6$, $b_2 = 18$ and so $N - b_1 = 24$, $N - b_2 = 12$. Thus, we seek the 18th and 19th largest elements of D_3. To find these we apply Case 4 of the Algorithm. We begin with

$$B(12,6) = \frac{X_{12}+X_{18}}{2} - \frac{X_6+X_{24}}{2} = 1.35 > 0.$$

Next we compute

$$B(11,7) = \frac{X_{11}+X_{17}}{2} - \frac{X_7+X_{25}}{2} = -0.17 < 0.$$

Thus, we have reached the stopping point and $\hat{\theta}(.2,.6)$ can be determined after two more comparisons.

Now $B(12,7) = .27 > 0$, $B(12,8) = -1.22 < 0$, $B(11,6) = .91 > 0$. Hence, the 7^{th} smallest element of D_{32} is between the 11^{th} and 12^{th} smallest in D_{31} and so it is 18^{th} smallest in D_3. Similarly, the 12^{th} smallest element of D_{31} is 19^{th} smallest in D_3. Thus

$$\hat{\Theta}(.2,.6) = \frac{1}{2}\left[\frac{X_7+X_{25}}{2} + \frac{X_{12}+X_{18}}{2}\right] = .79$$

Note that for this data $\overline{X} = .46$, the sample median is .78, and the median of the full set of Walsh averages is .705. The mean seems to have been greatly influenced by the observation -10.33. If we remove this point, the sample mean becomes .837, which is similar to the other estimators. Notice also that 5 Walsh averages and 5 comparisons were needed in the computations. This can be contrasted with the 465 Walsh averages which enter into computation of the median of the full set of Walsh averages, which is the estimator based on the Wilcoxon statistic.

For $k > 3$, the same technique can be applied although the number of ordered sets increases and as a result substantially more averages may be needed before the estimator can be determined.

Next, we consider the efficiency of $\hat{\theta}^k$. Let $\hat{\rho}_1$ and $\hat{\rho}_2$ be two consistent estimators of θ. Suppose that $N^{1/2}(\hat{\rho}_i - \theta)$ has a $N(0,\sigma^2(\rho_i))$ limiting distribution. Then the asymptotic relative efficiency of $\hat{\rho}_1$ with respect to $\hat{\rho}_2$ is $ARE(\hat{\rho}_1,\hat{\rho}_2) = \sigma^2(\hat{\rho}_2)/\sigma^2(\hat{\rho}_1)$. In addition, if $\hat{\rho}_i$ is the H-L estimator derived from a statistic R_N, then $\sigma^2(\hat{\rho}_i) = [C^2(R_N)]^{-1}$ and hence the ARE of two H-L estimators is the same as the Pitman efficiency of the corresponding test statistics. See Randles and Wolfe (1979, p. 227). In particular, since $T_{N,k}$ is asymptotically normal, $N^{1/2}(\hat{\theta}^k - \theta)$ has $N(0,[C^2(T_{N,k})]^{-1})$ as its limiting distribution, where $[C^2(T_{N,k})]^{-1}$ is given in (2.2).

As a result, the efficiency comparisons of Table 1-3 continue to hold when comparing the corresponding point estimators. In particular, using Table 3, we can compare $\hat{\theta}^3(.2,.6)$ with the sample mean, median, and the median of the set of all Walsh averages. We see that $\hat{\theta}^3(.2,.6)$ compares favorably with each of these estimators, and, in particular, its efficiency very closely parallels that of the median of all Walsh averages. However, computation of $\hat{\theta}^3(.2,.6)$ requires only a small subset of all Walsh averages, as we have previously seen. Thus, $\hat{\theta}^3(.2,.6)$ is a viable alternative to the other estimators if both simplicity and efficiency are considered.

Next, we propose an alternative estimator of θ which further utilizes the ordering of the elements of D_{31} and D_{32}. The estimator suggested is a linear combination of the medians of the sets D_{31} and D_{32} where the weights, λ, and $1 - \lambda$, are proportional to the number of elements in D_{31} and D_{32}, respectively. Following this approach, we obtain

$$\begin{aligned}\hat{\theta}^*(\gamma_1,\gamma_2) = (X_{[(1-\gamma_1)N/2]} + X_{[(1+\gamma_1)N/2]})\lambda/2 +& \\ (X_{[(1-\gamma_2)N/2]} + X_{[(1+\gamma_2)N/2]})(1 - \lambda)/2&\end{aligned} \tag{3.4}$$

where λ = proportion of elements in D_3 which are in D_{31}. Although $\hat{\theta}^*$ is not a H-L estimator based on any member of $T_{N,2}$, we would expect it to have similar properties. This is confirmed in Table 4 which compares $\hat{\theta}^*(.2,.6) = (X_{[.4N]} + X_{[.6N]})/3 + (X_{[.2N]} + X_{[.8N]})/6$ with $\hat{\theta}^3(.2,.6)$ as well as with other commonly used estimators of θ. We see that very little efficiency loss results from using $\hat{\theta}^*$ in place of $\hat{\theta}$, and some computational advantage is gained. Again note the improved efficiency of this simple estimator at the normal distribution. Thus, in terms of simplicity and efficiency, both $\hat{\theta}*(.2,.6)$ and $\hat{\theta}(.2,.6)$ are viable alternatives to the classical estimators considered in this study.

- Table 4 about here -

We conclude this section with a brief discussion of the favorable robustness properties of $\hat{\theta}^k$ and $\hat{\theta}*$. The influence curve of $\hat{\theta}^k$ can be determined using the methods of Hettmansperger and Utts (1977) and that of $\hat{\theta}*$ by following Huber (1977). In each case, we find that the influence curve is a step function with a finite number of jumps so that it is bounded but discontinuous. The boundedness is a desirable property since it indicates that the estimator is not unduly influenced by any single observation, regardless of its magnitude. The discontinuity of the influence indicates instability in the estimator at the points of discontinuity. A similar effect has been noted by Hampel (1974) for the Winsorized mean. This instability in $\hat{\theta}$ is further illustrated for the case k = 3 by the following result which gives a worst case comparison of $\hat{\theta}^3$ with $\overline{X}$.

Theorem 4. Let Ω be the class of symmetric, continuous, unimodal densities. Then

$$\inf_{f\varepsilon\Omega} \text{ARE}(\hat{\theta}(\gamma_1,\gamma_2), \overline{X}) = (1 - \gamma_1)^3/(3(1 - \gamma_2) + (\gamma_2 - \gamma_1)/4) \qquad (3.5)$$

We defer the proof until the appendix. Notice that when $\gamma_1 = .2$ and $\gamma_2 = .6$, the infimum is .39. This is a slight improvement over the median vs. the mean for which the infimum is .33. The infimum in the case of the median of the Walsh averages vs. the mean is .864. In addition, it should be noted that the least favorable distribution is one which exhibits maximum instability at those points which are discontinuities of the influence curve.

Another measure of robustness is the breakdown point (Huber 1981). While the influence curve measures the effect of an additional observation at y on the estimator, the breakdown point gives the fraction of gross errors the estimator can tolerate before it becomes unbounded. Hampel (1971) formulates a technical definition as well as some examples. The breakdown point of $\hat{\theta}^*(\gamma_1, \gamma_2)$ is $(1 - \gamma_2)/2$. This follows from Huber (1981) but is also intuitively clear since $(1 - \gamma_2)/2$ is the most extreme quantile used to compute $\hat{\theta}^*$. Following Hettmansperger and Utts (1977), it can be shown that the breakdown of $\hat{\theta}^3(\gamma_1, \gamma_2)$ is

$$b(\hat{\theta}^3(\gamma_1, \gamma_2)) = \begin{cases} (\gamma_2 - \gamma_1)/2 & 3\gamma_2 \geq \gamma_1 + 2 \\ 2^{-1} - (\gamma_1 + \gamma_2)/4 & 3\gamma_2 < \gamma_1 + 2 \end{cases}$$

Notice that the breakdown of $\hat{\theta}^*(.2,.6)$ is .2 while that of $\hat{\theta}(.2,.6)$ is .3. We also note that the breakdowns of $\bar{X}$, HL, and $\tilde{X}$ are 0, .293, and .50, respectively. Table 5 indicates how the breakdown of $\hat{\theta}(\gamma_1, \gamma_2)$ varies with γ_1 and γ_2.

- Table 5 about here -

4. Interval Estimation

Confidence intervals for θ based on rank tests are described in Lehmann (1963), Sen (1966), and Bauer (1972). Applying the methods of Bauer to the statistics $T_{N,k}$, we see that the endpoints, $\hat{\theta}_L^k$ and $\hat{\theta}_U^k$, of the confidence interval must be elements of D_k. In particular, if $P_{\theta=0}(T_{N,k} \leq C) = \alpha/2$, then a $(1 - \alpha)$ 100% confidence interval for θ has endpoints $\hat{\theta}_L^k$, the $(C + 1)$'st ordered element in D_k and $\hat{\theta}_U^k$, the $(N_k - C)$'th ordered element in D_k, where $N_k = kN - \sum_{i=1}^{k} b_i$, the number of elements in D_k. Thus, the problem of computing the endpoints of this interval is reduced to that of finding particular Walsh averages within the restricted set D_k.

If $k = 2$, we have previously seen that the elements of D_2 are ordered and so we have $\hat{\theta}_L^2 = (X_{C+1} + X_{C+1+b_1})/2$, and $\hat{\theta}_U^2 = (X_{N-b_1-C} + X_{N-C})/2$ where C can be determined using the fact that $T_{N,2}(0) \sim \text{Bin}(N - b_1, 1/2)$ or by a normal approximation. This result was first obtained by Noether (1973). If $k = 3$, we seek the $(C + 1)$'st and $(2N - b_1 - b_2 - C)$'th ordered Walsh averages in D_3. The algorithm of Section 3 modified to start from selected quantiles of the sets D_{31} and D_{32} can be used to compute these endpoints. We illustrate this with the data previously considered. We will calculate an approximate 95.44% confidence interval for θ. Using the Corollary to Theorem 1, a normal approximation with continuity correction yields

$$C \doteq \Phi^{-1}(\alpha/2)((b_2 - b_1)/4 + N - b_2)^{1/2} + (b_2 - b_1)/2 + N - b_2 - 1/2. \tag{4.1}$$

Now letting $\gamma_1 = .2$, $\gamma_2 = .6$ and recalling that $N = 30$, we obtain $C \doteq 10$. Thus, $\hat{\theta}_L$ is the 11'th smallest element in D_3 and $\hat{\theta}_U$ is the 26'th smallest element in D_3. To compute $\hat{\theta}_L$, notice that the number of Walsh averages in D_{31} is 12. As a result, we first consider B(7,4)(since 11(24)/36 $\doteq$ 7 and 11(12)/36 $\doteq$ 4). In applying this method, care is needed to assure that the starting point be of the form B(a,b) with a + b = d if the d'th smallest element is desired. Now, B(7,4) < 0 so we check B(8,3) > 0. Furthermore, B(8,4) > 0 and B(7,3) > 0. Thus, the 7'th element of D_{31} is 11'th smallest, i.e. $\hat{\theta}_L = (X_7 + X_{13})/2 = -1.41$. By a similar argument starting with B(17,9) we find B(17,9) > 0, B(16,10) < 0, B(17,10) > 0, and B(16,9) > 0. Thus, $\hat{\theta}_U = (X_{16} + X_{22})/2 = 2.36$. Notice that only 8 Walsh averages were required in this case, whereas there are a total of 465 Walsh averages for this sample. The approximate 95.44% confidence interval based on the Wilcoxon statistic which requires ordering of these 465 Walsh averages is (-1.04, 2.30) for this data.

Lehmann (1963) has considered two definitions of the asymptotic relative efficiency of two interval estimation procedures; one in terms of the probability of covering a false value, the other in terms of the asymptotic behavior of the lengths. Randles and Wolfe (1979) denote these by ARE and L-ARE, respectively. Lehmann has shown that in many regular cases, including all those considered in this work, the two definitions yield the same efficiencies. Furthermore, the efficiencies of confidence intervals derived from test statistics using the method of Lehmann is just the ratio of squared efficacies. Hence, we see that the favorable efficiencies of $T_{N,k}$ and $\hat{\theta}^k$ also apply to the interval $(\hat{\theta}^k_L, \hat{\theta}^k_U)$. In particular, if $\gamma_1 = .2$, $\gamma_2 = .6$, the interval $(\hat{\theta}_L{}^3, \hat{\theta}_U{}^3)$ has efficiency similar to the interval derived from the Wilcoxon statistic, yet requires computation of only a very small number of Walsh averages.

Next, we consider another confidence interval for θ obtained by determining two separate intervals, one from D_{31} and the other from D_{32}, and then combining the endpoints of these intervals. Suppose that the desired confidence coefficient is $1 - \alpha$. Then an approximate $(1 - \alpha)$ 100% confidence interval for θ based on the elements of D_{31} will be $(W_{C_1+1}, W_{N-b_1-C_1})$ where $W_1 \leq W_2 \leq \ldots \leq W_{N-b_1}$ are the ordered elements of D_{31} and using a normal approximation $C_1 \doteq \Phi^{-1}(\alpha/2)\ (N - b_1)^{1/2}\ /2 + (N - b_1 - 1)\ /2$. Similarly, an approximate $(1 - \alpha)$ 100% confidence interval for θ based on the elements of D_{32} is $(Z_{C_2+1}, Z_{N-b_2-C_2})$ where $Z_1 \leq \ldots \leq Z_{N-b_2}$ are the ordered elements of D_{32} and $C_2 \doteq \Phi^{-1}\ (\alpha/2)\ (N - b_2)^{1/2}/2 + (N - b_2 - 1)/2$. Now let

$$\hat{\theta}_L{}^* = \lambda\, W_{C_1+1} + (1 - \lambda)\, Z_{C_2+1}$$

$$\hat{\theta}_U{}^* = \lambda\, W_{N-C_1-b_1} + (1 - \lambda)\, Z_{N-C_2-b_2} \tag{4.2}$$

where $0 \leq \lambda \leq 1$. Then we propose $(\hat{\theta}_L{}^*, \hat{\theta}_U{}^*)$ as a confidence interval for θ.

<u>Theorem 5</u>. $(\hat{\theta}_L{}^*, \hat{\theta}_U{}^*)$ is asymptotically a $(1 - \alpha)$ 100% confidence interval for θ; ie, $\lim\limits_{N\to\infty} P(\hat{\theta}_L{}^* < \theta < \hat{\theta}_U{}^*) = 1 - \alpha$.

<u>Proof</u>: As N tends to infinity, we have $P(\theta < W_{C_1+1})$, $P(\theta < Z_{C_2+1})$, $P(\theta > W_{N-b_1-C_1})$ and $P(\theta > Z_{N-b_2-C_2})$ each tending to $\alpha/2$. But

$\min\ \{W_{C_1+1}, Z_{C_2+1}\} \leq \hat{\theta}_L{}^* \leq \max\ \{W_{C_1+1}, Z_{C_2+1}\}$ and

$\min\ \{W_{N-b_1-C_1}, Z_{N-b_2-C_2}\} \leq \hat{\theta}_U^* \leq \max\ \{W_{N-b_1-C_2}, Z_{N-b_2-C_2}\}$.

Combining these results implies that $P(\theta < \hat{\theta}_L{}^*)$ tends to $\alpha/2$ and $P(\theta > \hat{\theta}_U{}^*)$

tends to $\alpha/2$ which proves the theorem.

Next, we investigate the efficiency of this interval with $\lambda = (1 - \gamma_1) / (2 - \gamma_1 - \gamma_2)$, so that the weights are proportional to the number of elements in the sets D_{31} and D_{32}. Notice that, in this case,

$$\hat{\theta}_L^* = \lambda\ (X_{C_1+1} + X_{C_1+b_1+1}) / 2 + (1 - \lambda)\ (X_{C_2+1} + X_{C_2+b_2+1})/2$$

$$\hat{\theta}_U^* = \lambda\ (X_{N-b_1-C_1} + X_{N-C_1}) / 2 + (1 - \lambda)\ (X_{N-b_2-C_2} + X_{N-C_2}) / 2 \tag{4.3}$$

Thus, $(\hat{\theta}_U^* - \hat{\theta}_L^*)$ is just a linear combination of the lengths of the confidence intervals based on D_{31} and D_{32}, respectively. As a result, we can use earlier results on efficiencies to compute the efficiency of the interval $(\hat{\theta}_L^*, \hat{\theta}_U^*)$. In particular, it can be shown that

$$[N^{1/2} (\frac{\hat{\theta}_U^* - \hat{\theta}_L^*}{2\ \Phi^{-1}\ (\gamma/2)})\ ]^2 \rightarrow \{\frac{\lambda}{C(T_{N,1}(\gamma_1)} + \frac{1 - \lambda}{C(T_{N,1}(\gamma_2)}\}^2 \tag{4.4}$$

in probability.

Using (4.4), Table 6 compares this interval with other commonly used confidence intervals for θ as well as with $(\hat{\theta}_L, \hat{\theta}_U)$ with $\gamma_1 = .2$, $\gamma_2 = .6$. We see that considerable efficiency is lost for both the logistic and double exponential distributions when $(\hat{\theta}_L^*, \hat{\theta}_U^*)$ replaces $(\hat{\theta}_L, \hat{\theta}_U)$. This is in contrast to the estimation case where the efficiency of $\hat{\theta}^*$ closely paralleled that of $\hat{\theta}$. Thus, we see that in the case of interval estimation, a somewhat higher price in terms of efficiency must be paid to get the additional simplicity of $(\hat{\theta}_L^*, \hat{\theta}_U^*)$ over $(\hat{\theta}_L, \hat{\theta}_U)$. Once again, however, both of these methods perform quite well at the normal model, and can be superior to the interval based on the sample mean for non-normal models.

- Table 6 about here -

5. Extensions and Conclusions

First, we briefly summarize, without proofs, the extensions of the previous results to the two-sample location problem. In particular, we point out that the scores function as a step function produces a two-sample rank statistic and corresponding point and interval estimators with the same asymptotic relative efficiencies as those considered in the one-sample problem. For a complete discussion, see Markowski (1980).

Let $X_1 < \ldots < X_m$ and $Y_1 < \ldots < Y_n$ be the ordered values of random samples from distributions with distribution functions $F(x)$ and $F(y - \delta)$, respectively. Suppose without loss of generality $m \leq n$.

Let
$$c_k(i) = \begin{cases} k & i = 0, \ldots, [(1-\gamma_k)\ N/2] \\ \cdot & \\ \cdot & \\ 0 & i = [(1-\gamma_1)\ N/2] + 1, \ldots, [(1+\gamma_1)\ N/2] \\ \cdot & \\ \cdot & \\ k & i = [(1+\gamma_k)\ N/2] + 1, \ldots, N \end{cases} \tag{5.1}$$

where $0 \leq \gamma_1 < \gamma_2 < \ldots < \gamma_k \leq 1$; $k < N = m + n$, k fixed.

Consider $S_{N,k}(\delta) = \sum_{i=1}^{n} c_k\ (R_i(\delta))$ where $R_i(\delta)$ is the rank of $Y_i - \delta$ among $X_1, \ldots, X_m, Y_1 - \delta, \ldots, Y_n - \delta$. Now with $\gamma_0 = 0$ and $\gamma_{k+1} = 1$ let

$$g_t = [(1 + \gamma_t)\ N/2],\quad g_{-t} = [(1 - \gamma_t)\ N/2] \quad t = 0, \ldots, k + 1. \tag{5.2}$$

<u>Theorem 6</u>. If $\delta = 0$, then $S_{N,\ k}(0) = \sum_{i=-k}^{k} i\ W_i$ where

$P(W_i = t) = \binom{g_i - g_{i-1}}{t} \binom{N - (g_i - g_{i-1})}{n-t} / \binom{N}{n}$ for $t = 0, 1, \ldots,$ $\min\{n, g_i - g_{i-1}\}$. In addition, $E(S_{N,\ k}(0)) = 0$, $\mathrm{Var}\ (S_{N,\ k}(0)) =$

$2\,mn\,(\sum_{i=1}^{k} i^2\,(g_{i+1} - g_i))\;/\;N(N - 1)$, and for k fixed, the limiting distribution (as $N \to \infty$) of $S_{N,\,k}(0)\;/\;(\mathrm{Var}\;S_{N,\,k}(0))^{1/2}$ is $N(0,\,1)$.

Next, we consider the H-L estimator derived from $S_{N,\,k}(\delta)$, which will be denoted by $\hat{\delta}_k$.

Theorem 7. Let g_t be as in (5.2), and $L_k = \{Y_j - X_i : i + j = g_t + 1,$ $t = \pm 1, \ldots \pm k\}$. Then $\hat{\delta}_k$ = med L_k. In particular,

$$\hat{\delta}_1 = \text{med } L_1 = \text{med } \{P_1 \cup P_{-1}\} \tag{5.3}$$

$$\hat{\delta}_2 = \text{med } L_2 = \text{med } \{P_1 \cup P_{-1} \cup P_{-2} \cup P_2\}$$

where $P_t = \{Y_j - X_i = i + j = g_t + 1\}$.

Notice that the elements within any set P_t are automatically ordered and so computation of $\hat{\delta}_1$ required the same steps as computation of $\hat{\theta}_2$ for which an algorithm was previously given. Computation of $\hat{\delta}_2$ requires the determination of the median of the union of four ordered sets. The algorithm considered here can be extended in a straightforward way to allow simple computation of $\hat{\delta}_2$.

In the same way, it can be shown that a $(1 - \alpha)$ 100% confidence interval for δ has the form $(\hat{\delta}_L, \hat{\delta}_U)$ where $\hat{\delta}_L$ = (C + 1) 'st smallest element in L_k

$$\hat{\delta}_U = (N_k' - C) \text{ 'th smallest element in } L_k \tag{5.4}$$

N_k' = the number of elements in L_k and C satisfies $P_{\delta=0}(S_{N,\,k}(0) \le C) = \alpha/2$. Thus, we see that computation of this interval requires the same type of algorithm as is needed to compute the estimator. Also, approximate confidence intervals can be obtained using the normal approximation given in Theorem 6.

In addition, all the efficiency comparisons in the one-sample case remain valid when comparing these inference methods with other two-sample methods. Also, the simplicity of the methods remains in the two-sample problem. Thus, once again we have proposed a class of inference methods which are desirable if efficiency, simplicity, and robustness are considered.

Lastly, the scores function considered can be adapted to the problem of making inferences about the slope parameter in a simple linear regression model when the form of the error distribution is unspecified. Once again, the resulting inference procedures will retain the efficiencies of those which we have discussed.

REFERENCES

Albers, W., Bickel, P. J., and van Zwet, W. R. (1976). Asymptotic expansions for the power of distribution free tests in the one-sample problem. The Annals of Statistics, 4, 1, 108-156.

Bauer, D. F. (1972). Constructing confidence sets using rank statistics. J. Amer. Statist. Assn., 67, 687-690.

Hampel, F. R. (1971). A general qualitative definition of robustness. Ann. Math. Statist., 42, 1887-1896.

Hettmansperger, T. P. and Utts, J. M. (1977). Robustness properties for a simple class of rank estimates. Commun. Statist., A6(9), 855-868.

Hodges, J. L., Jr. and Lehmann, E. L. (1963). Rank estimates of location. Ann. Math. Statist., 34, 598-611.

Huber, P. J. (1981). Robust Statistics. John Wiley & Sons, New York.

Johnson, D. B. and Ryan, T. A., Jr. (1978). Fast computation of the Hodges-Lehmann estimator - Theory and practice. 1978 Proc. Statist. Assn. Ann. Meeting, San Diego, California. 2 pp.

Lehmann, E. L. (1963). Nonparametric confidence intervals for a shift parameter. Ann. Math. Statist., 34, 1507-1512.

Lehmann, E. L. (1975). Nonparametrics: Statistical Methods Based on Ranks. San Francisco: Holden Day.

Markowski, E. P. (1980). Simple Estimation Procedures Based on Rank Statistics for the One- and Two- Sample Location Models. Unpublished Dissertation, The Pennsylvania State University.

Noether, G. E. (1973). Some simple distribution-free confidence intervals for the center of a symmetric distribution. J. Amer. Statist. Assn., 68, 716-719.

Policello, G. E., II and Hettmansperger, T. P. (1976). Adaptive robust procedures for the one-sample location problem. J. Amer. Statist. Assn., 71, 624-633.

Randles, R. H. and Wolfe, D. A. (1979). Introduction to the Theory of Nonparametric Statistics. John Wiley & Sons, New York.

APPENDIX

Lemma 1. Let Ω be the class of symmetric, unimodal densities, then

$$\inf_{f\varepsilon\Omega} e(\hat{\theta}(\gamma_1), \overline{X}) = \frac{1}{3}(1-\gamma_1)^2. \tag{A.1}$$

Proof:

$$e(\hat{\theta}(\gamma_1), \overline{X}) = \frac{4f^2[F^{-1}(\frac{1+\gamma_1}{2})]}{1-\gamma_1} \sigma_f^2 . \tag{A.2}$$

We assume WLOG that $f(0) = 1$. We wish to find $f \varepsilon \Omega$ which minimizes $f[F^{-1}(\frac{1+\gamma_1}{2})] \sigma_f^2$ for fixed γ_1. Let

$$f_c(x) = \begin{cases} 1 & 0 \leq x < \gamma_1/2 \\ \frac{1-\gamma_1}{2c} & \gamma_1/2 \leq x \leq c + \gamma_1/2 \end{cases} \tag{A.3}$$

and $f_c(-x) = f_c(x)$.

Then it is easily verified that f is a density function and it is clear that f may be approximated arbitrarily closely by a density f_1 in Ω.

Furthermore, $f_c[F_c^{-1}(\frac{1+\gamma_1}{2})] = \frac{1-\gamma_1}{2c}$ and $\sigma_{f_c}^2 = \frac{2}{3}(\frac{\gamma_1}{2})^3 + \frac{1-\gamma_1}{3}c^2 + \frac{\gamma_1(1-\gamma_1)}{2}c + \frac{\gamma_1{}^2(1-\gamma_1)}{4}$.

Now for $f[F^{-1}(\frac{1+\gamma_1}{2})]$ fixed, σ_f^2 is minimized for unimodal densities by choosing f to be uniform for $x \geq \gamma_1/2$. Hence, the density f_c will be made least favorable by letting $c \to \infty$. In this case,

$$e(\hat{\theta}(\gamma_1),\overline{X}) = \frac{4}{1-\gamma_1}\left(\frac{1-\gamma_1}{2c}\right)^2 \sigma_f^2$$

and letting $c \to \infty$ we have

$$\inf_{f\varepsilon\Omega} e(\hat{\theta}(\gamma_1),\overline{X}) = \frac{1-\gamma_1}{3}\left(\frac{4}{1-\gamma_1}\right)\left(\frac{1-\gamma_1}{2}\right)^2 = \frac{1}{3}(1-\gamma_1)^2 .$$

Finally, we note that $f_c \notin F_u$ but since it can be approximated arbitrarily closely by $f_1 \varepsilon F_u$, the result remains valid.

Notice that the least favorable density is 1 until the point $\frac{\gamma_1}{2}$ and then becomes arbitrarily small and uniform in the tails. We now turn to the result comparing $\hat{\theta}(\gamma_1,\gamma_2)$ and $\overline{X}$. First, we need a preliminary lemma.

Lemma 2.

$$\lim_{c\to\infty} e_{f_c}(\hat{\theta}(\gamma_1,\gamma_2),\overline{X}) \leq e_f(\hat{\theta}(\gamma_1,\gamma_2),\overline{X})$$

for any $f \varepsilon F_u$ where f_c is given by A.3.

Proof: Again take $f(0) = 1$. It is clear from the proof of Lemma 1 that the least favorable density must be of the form:

$$g(x) = \begin{cases} 1 & 0 \le x < \gamma_1/2 \\ a & \gamma_1/2 \le x < \gamma_2/2 \\ b & \gamma_2/2 \le x < \gamma_2/2 + d \end{cases}$$

where $b \le a \le 1$ and $d \ge 0$.

But by the previous lemma, for fixed a, $e(\hat{\theta}(\gamma_1,\gamma_2),\bar{X})$ is minimized by letting $b \to 0$. Furthermore, for any fixed a and b, we may decrease the efficiency by choosing $a' < a$ and letting $b \to 0$. Thus, the least favorable choice for g makes a arbitrarily small and $b = 0$. But this is just the density f_c with c becoming arbitrarily large, proving the result.

Proof of Theorem 4.

By Lemma 2, $\inf_{f\varepsilon\Omega} e(\hat{\theta}(\gamma_1,\gamma_2),\bar{X}) = \lim_{c\to\infty} e_{f_c}(\hat{\theta}(\gamma_1,\gamma_2),\bar{X})$

$$= \lim_{c\to\infty} \frac{(\frac{1-\gamma_1}{2c} + \frac{1-\gamma_1}{2c})^2}{1-\gamma_2 + \frac{1}{4}(\gamma_2-\gamma_1)} \sigma^2_{f_c}$$

$$= \frac{(1-\gamma_1)^2}{1-\gamma_2 + \frac{1}{4}(\gamma_2-\gamma_1)} \cdot \frac{(1-\gamma_1)}{3} = \frac{(1-\gamma_1)^3}{3(1-\gamma_2 + \frac{1}{4}(\gamma_2-\gamma_1))} .$$

Table 1 Maximum Efficacy of $T_{N,\ 1}$

	Kurtosis	C^{2*}	γ_1^*	$e(T_{N,1},\ LRT)$
A_1	3	.81	.46	.81
A_2	4.2	.296	1/3	.89
A_3	7.65	.74	.38	.84
A_4	6	1	0	1

C^{2*} is maximum squared efficacy of $T_{N,\ 1}$

γ_1^* is value of γ_1 which maximizes C^*

$e(T_{N,\ 1,}\ LRT)$ is efficiency of $T_{N,\ 1}$ with $\gamma_1 = \gamma_1^*$ to likelihood ratio test (LRT)

2. Efficacy of $T_{N,2}$ as a function of γ_1, γ_2 and F

C^2

γ_1	γ_2	A_1	A_2	A_3	A_4
0	0	.637	.250	.595	1*
.1	.2	.742	.288	.692	.876
.1	.4	.819	.310	.757	.833
.1	.6	.870	.316	.795	.805
.1	.8	.870	.304	.780	.807
.2	.4	.829	.312	.765	.865
.2	.6	.888	.320*	.808*	.720
.2	.8	.902	.311	.807	.714
.4	.6	.876	.304	.789	.556
.4	.8	.912*	.300	.803	.533
.4	.9	.903	.295	.764	.544
.5	.7	.866	.283	.775	.457
.5	.8	.884	.280	.774	.445
.5	.9	.885	.276	.746	.450
.6	.8	.830	.250	.713	.360
.6	.9	.839	.246	.688	.357
.7	.9	.754	.204	.604	.267
.8	.9	.620	.151	.467	.180

*denotes optimal choice of (γ_1, γ_2).

3. Efficiency of $T_{N,2}$ with $\gamma_1 = .2, \gamma_2 = .6$

	A_1	A_2	A_3	A_4
$e(T_{N,2}, S)$	1.40	1.28	1.32	.72
$e(T_{N,2}, W)$	.930	.96	.99	.96
$e(T_{N,2}, t)$	.890	1.05	1.13	1.44
$e(T_{N,2}, T^*_{N,2})$	.976	1	1	.72
$e(T_{N,2}, LRT)$	.890	.961	.960	.72

S - sign test, W - Wilcoxon, t - t-test, $T_{N,2}^*$ - optimal choice of (γ_1, γ_2)

Table 4

ARE of $\hat{\theta}^*(.2, .6)$ with respect to other estimators of θ.

	A_1	A_2	A_3	A_4
ARE($\hat{\theta}^*$, $\bar{X}$)	.87	1.04	1.12	1.44
ARE($\hat{\theta}^*$, HL)	.91	.96	.98	.96
ARE($\hat{\theta}^*$, $\tilde{X}$)	1.37	1.28	1.30	.72
ARE($\hat{\theta}^*$, $\hat{\theta}(.2, .6)$)	.98	.99	.99	1.00

Note: $\bar{X}$ - sample mean, HL - median of Walsh averages, $\tilde{X}$ - sample median

5. Breakdown of $\hat{\theta}(\gamma_1, \gamma_2)$

γ_1	0	.1	.1	.1	.1	.2	.2	.2	.4	.4
γ_2	0	.2	.4	.6	.8	.4	.6	.8	.6	.8
ε^*	.500	.425	.375	.325	.350	.350	.300	.300	.250	.200

γ_1	.4	.5	.5	.5	.6	.6	.7	.8
γ_2	.9	.7	.8	.9	.8	.9	.9	.9
ε^*	.250	.200	.175	.200	.150	.150	.100	.075

6. ARE of $(\hat{\theta}_L^*, \hat{\theta}_U^*)$ relative to other interval estimators of θ, $(\gamma_1 = .2, \gamma_2 = .6)$

	A_1	A_2	A_3	A_4
ARE(1, 2)	.87	.91	1.10	1.24
ARE(1, 3)	1.37	1.11	1.28	.62
ARE(1, 4)	.91	.84	.96	.83
ARE(1, 5)	.98	.87	.97	.86

1 - $(\hat{\theta}_L^*(.2, .6), \hat{\theta}_U^*(.2, .6))$, 2 - t-interval, 3 - sign test interval

4 - Wilcoxon test interval, 5 - $(\hat{\theta}_L(.2, .6), \hat{\theta}_U(.2, .6))$

Unclassified

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

REPORT DOCUMENTATION PAGE

READ INSTRUCTIONS BEFORE COMPLETING FORM

1. REPORT NUMBER Penn State Tech. Rpt. #37	2. GOVT ACCESSION NO. AD-A101 734	3. RECIPIENT'S CATALOG NUMBER
4. TITLE (and Subtitle) Inference Based on Simple Step Statistics for the Location Model		5. TYPE OF REPORT & PERIOD COVERED 6. PERFORMING ORG. REPORT NUMBER
7. AUTHOR(s) Edward P. Markowski, Old Dominion University Thomas P. Hettmansperger, Penn State University		8. CONTRACT OR GRANT NUMBER(s) N00014-80-C-0741
9. PERFORMING ORGANIZATION NAME AND ADDRESS Department of Statistics Penn State University University Park, PA 16802		10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS NR 042-446
11. CONTROLLING OFFICE NAME AND ADDRESS Office of Naval Research Statistics and Probability Program Code 436 Arlington, VA 22217		12. REPORT DATE July 1981 13. NUMBER OF PAGES 34
14. MONITORING AGENCY NAME & ADDRESS(if different from Controlling Office)		15. SECURITY CLASS. (of this report) Unclassified 15a. DECLASSIFICATION/DOWNGRADING SCHEDULE

16. DISTRIBUTION STATEMENT (of this Report)

APPROVED FOR PUBLIC RELEASE: DISTRIBUTION UNLIMITED.

17. DISTRIBUTION STATEMENT (of the abstract entered in Block 20, if different from Report)

18. SUPPLEMENTARY NOTES

19. KEY WORDS (Continue on reverse side if necessary and identify by block number)

Robust confidence intervals, Robust tests, rank tests, rank estimates

20. ABSTRACT (Continue on reverse side if necessary and identify by block number)

Until recent years, statistical theory has emphasized optimal procedures under an assumed underlying model. However, these methods may be highly sensitive to slight departures from the model assumptions (non-robust), and they may be inferior to other procedures for these alternative models (inefficient). The sample mean is an example of an estimator which suffers from these criticisms.

(continued)

Unclassified

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

In this paper, a class of estimation procedures which are robust, relatively efficient, and yet computationally simple is proposed for the one- and two-sample location problems. Particular emphasis is placed on the problem of determining confidence intervals with the properties stated above, a topic which has seen limited exposure in the literature. As a by-product, some interesting point estimates are obtained.

A class of score functions which are ordinary step functions is considered for the location model. Point estimates and confidence intervals are obtained by inverting the corresponding rank statistics. Efficiency and robustness properties of the procedures are investigated. Several computational schemes are illustrated which make the estimates and confidence intervals quite easy to compute.

Unclassified

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)